计 算 机 科 学 丛 书

数据挖掘
概念与技术

Data Mining
Concepts and Techniques

(加) Jiawei Han
Micheline Kamber 著
范 明 孟小峰 等译

机械工业出版社
China Machine Press
MORGAN KAUFMANN

计算机科学丛书

数据挖掘

概念与技术

（加）Jiawei Han
Micheline Kamber 著

范 明 孟小峰 等译

机械工业出版社
China Machine Press

数据挖掘是数据库研究、开发和应用最活跃的分支之一。本书从数据库角度全面、系统地介绍数据挖掘的基本概念、基本方法和基本技术以及数据挖掘的最新进展，是一本可读性极佳的教材。

本书全面而深入地叙述了数据库技术的发展和数据挖掘应用的重要性，数据仓库和OLAP（联机分析处理）技术，数据预处理技术（包括数据清理、数据集成和转换、数据归约的方法），数据挖掘技术（包括分类、预测、关联和聚类等基础概念和技术），先进的数据库系统中的数据挖掘方法，数据挖掘的应用和一些具有挑战性的研究问题。作者注重实效，将以上内容辅以实例，对每类问题均提供代表性算法，并给出每一技术具体的应用法则。该书由10章及两个附录组成。通过本书的学习，读者可以对数据挖掘的整体结构、概念和技术有深入的认识和了解，并且可以熟悉数据挖掘的基本原理和发展方向。

本书适合作为相关专业高年级本科生的选修课教材，特别适合作为研究生的专业课教材，同时也可供从事数据挖掘研究和应用开发工作的相关人员作为必备的参考书。

本书版权登记号：图字：01-2001-2213

图书在版编目（CIP）数据

数据挖掘：概念与技术/（加）韩家炜，（加）坎伯（Kamber, M.）著；范明等译. - 北京：机械工业出版社，2001.8
（计算机科学丛书）
书名原文：Data Mining: Concepts and Techniques
ISBN 7-111-09048-9

Ⅰ. 数... Ⅱ. ①韩... ②坎... ③范... Ⅲ. 数据处理 Ⅳ. TP274

中国版本图书馆CIP数据核字（2001）第039448号

机械工业出版社（北京市西城区百万庄大街22号 邮政编码 100037）
责任编辑：李伯民
北京牛山世兴印刷厂印刷 · 新华书店北京发行所发行
2001年8月第1版第1次印刷
787mm × 1092mm 1/16 · 24.5印张
印数：0 001 - 5 000册
定价：39.00元

译 者 序

数据挖掘是数据库研究、开发和应用最活跃的分支之一。这是很自然的事。数据库系统特别是关系数据库系统的成功，使我们有了强有力的事务处理工具。在计算机的帮助下，人们可以将传统的事务处理做得更好。不满足现状是社会前进的动力。人类当然不会满足于让计算机仅仅做事务处理。试图将数据库技术应用到更广泛的领域，导致了对时间数据库、空间数据库、多媒体数据库、工程数据库、统计数据库等面向特殊应用的数据库系统的研究与开发。新的应用导致对新的数据模型的需求，从而激发了扩充关系的、面向对象的、对象－关系的、演绎的等新数据模型和数据库系统的研究和开发。各种各样的数据库系统的开发，使得更多的数据以前所未有的速度收集在计算机中。人们当然不会仅仅满足对这些数据的简单查询。从信息处理的角度，人们更希望计算机帮助我们分析数据、理解数据，帮助我们基于丰富的数据作出决策，做人力所不能及的事情。于是，数据挖掘——从大量数据中用非平凡的方法发现有用的知识——就成了一种自然的需求。正是这种需求引起了人们的广泛关注，导致了数据挖掘研究的蓬勃开展。

数据挖掘是一个多学科交叉领域。这同样是很自然的事。一方面，想要以非平凡的方法发现蕴藏在大量数据集中的有用知识，数据挖掘必须从数据库技术、人工智能、机器学习、神经网络、统计学、模式识别、知识库系统、知识获取、信息提取、高性能计算和数据可视化等学科领域汲取营养。另一方面，这些学科领域也要发展，也在从不同角度关注数据的分析和理解，数据挖掘也为这些学科领域的发展提供了新的机遇与挑战。

数据挖掘引起了学术界和产业界的广泛关注，吸引了一大批研究者和开发者。国内外许多大学都先后开设了数据挖掘课程。然而，长期以来并没有合适的教材或专著。1999年9月，在美国San Diego的KDD99国际会议上，我们得知Jiawei Han（韩家炜）教授和Micheline Kamber正在写一本关于数据挖掘的书。不久，我们得到了韩家炜的《数据挖掘：概念和技术》书稿的前8章和第9、10两章的目录。浏览了各章目录并认真地阅读几章后，我们被这本书深深地吸引了。在此之前，译者看过几本关于数据挖掘的书。就译者所知，从数据库角度全面、系统地介绍数据挖掘的基本概念、基本方法和基本技术以及数据挖掘的最新进展，《数据挖掘：概念和技术》还是第一本。这使译者萌发了将该书译成中文，介绍给国内同行的念头。

Jiawei Han（韩家炜）教授是数据库领域国际知名的学者。他早年就读于郑州大学，后赴美国留学，在威斯康辛大学获硕士和博士学位。毕业后，他曾在美国西北大学任教，1988年起在加拿大西蒙·弗雷泽大学任教，现任计算科学系教授、智能数据库系统研究实验室主任。他是KDD等十几个国际学术会议的程序委员会委员，《IEEE知识与数据工程汇刊》、《数据挖掘与知识发现》等多种学术期刊的编委。韩家炜教授在演绎数据库、数据挖掘、数据库系统等方面的研究一直居领先地位。他先后在国际著名学术刊物和重要国际学术会议上发表论文100余篇，主持开发了数据挖掘系统DBMiner。《数据挖掘：概念和技术》建立了一个学习数据挖掘的有组织的框架，也融入了韩家炜教授从事数据挖掘研究十余年的心血。

正如Jim Gray所指出的，数据挖掘领域“发展非常迅速，这本书提供了一条学习该领域基本思想和了解该领域现状的快捷之路。”

全书主要包括10章和两个附录。本书的翻译和审校由范明和孟小峰共同组织完成。范明负责第1～7章。孟小峰负责第8～10章及两个附录。参加翻译工作的还有徐华（第1、2章），叶阳东（第3、4章），姬安明（第7章），王静（第8章），李盛恩（附录A），李翠萍（附录B）。此外，北京石油大学马玉书教授审阅了译稿的第1～7章，提出了许多宝贵的意见和建议；北京大学计算机系董云海对第8章提出了修改意见。全书由范明和孟小峰负责统一定稿。译者还参照该书Web主页中的勘误表，对书中的疏漏之处进行了更正。同时对在翻译中发现的错误进行了更正，并反馈给作者。

在本书翻译过程中，得到韩家炜教授的大力支持。他向译者提供了书稿第1～8章和英文版的最终版本。译者感谢机械工业出版社华章公司的编辑们，是他们的远见，使得该书能够尽快与读者见面。

由于本书涉及面广，许多术语目前尚无固定译法，翻译难度确实很大。有时，为了对一个术语选择一个简洁、达义的译法，译者虽经反复推敲、讨论，但仍然难免出现词不达意之处。此外，由于译者水平有限，译文中的不当之处也在所难免。译文中的错误当然应当由译者负责。但我们真诚地希望同行和读者朋友们不吝赐教。如果你能将你的意见和建议发往mfan@mail.zzu.edu.cn或xfmeng@public.bpa.net.cn，我们将不胜感激。

译　者

2001年4月

译者简介

范明　郑州大学计算机科学系教授，副系主任，兼任河南省计算机学会软件专业委员会主任。长期从事计算机软件教学和研究。主要讲授的课程包括计算机操作系统、数据库系统原理、知识库系统原理、数据挖掘和程序设计等。关心的主要研究领域包括递归查询优化、数据挖掘和数据仓库。1989—1990年曾访问加拿大西蒙·弗雷泽大学计算机科学系，从事演绎数据库研究。1999年访问美国Wright state大学计算机科学与工程系、从事数据挖掘研究。曾与南京大学徐洁磐教授合作主持国家自然科学基金项目1项，主持河南省自然科学基金和科技攻关项目多项。近十年发表论文20余篇，参加了《数据库综合大辞典》(1995)和《数据库技术新进展》(1997)的编写，并与徐洁磐。马玉书合作出版著作《知识库系统导论》(2000)。

孟小峰　博士，教授，中国人民大学信息学院计算机系副主任，中国计算机学会理事，中国计算机学会数据库专业委员会委员、秘书长，中国计算机学会青年计算机科技论坛(YOCSEF)副主席，多次担任国际学术会议程序委员会委员，目前为《计算机研究与发展》编委。1994—1996年曾在香港中文大学和城市大学从事研究工作。主持或参加过十多项国家科技攻关项目、国家自然科学基金以及国家863项目，获国家科技进步二等奖，电子部科技进步特等奖，北京市科技进步二等奖等奖励。研制开发的主要软件产品有国产数据库系统COBASE、嵌入式移动数据库系统“小精灵”、中文自然语言查询系统NChiql和并行数据库系统PBASE／1等。近十年在国内外杂志及国际会议发表论文50余篇，有数据库方面的著译作七部。主要研究领域为数据库系统实现技术、数据库查询语言、自然语言接口、嵌入与移动数据管理、web数据管理等。

序

我们被数据——科学数据、医疗数据、人口统计数据、财经数据和市场数据——淹没。人们没有时间看数据。人类的关注已经成为一种宝贵的资源。因此，我们必须找到有关方法，自动地分析数据、自动地对数据分类、自动地对数据汇总、自动地发现和描述数据中的趋势、自动地标记异常。这是数据库研究最活跃、最令人激动的领域之一。诸如统计、可视化、人工智能和机器学习方面的研究者正在为该领域做出贡献。该领域的宽广使得很难把握它过去几年的非凡进展。

Jiawei Han和Micheline Kamber做了一件极好的工作，在这本可读性极佳的教材中组织和提供了数据挖掘的内容。他们从介绍数据库和数据挖掘概念入手，特别强调了数据分析的需求。通过提供一个一般框架，综述了当前产品的情况。然后，逐章介绍了分类、预测、关联和聚类等基础概念和技术。作者注重实效，将这些内容辅以实例，对每类问题均提供代表性算法，并给出每一技术具体应用的经验法则。我认为这种写作风格具有很好的可读性，并且我已通过阅读该书学到了许多。Jiawei Han和Micheline Kamber在数据挖掘研究方面一直处于领先地位。这是一本他们用于培养自己的学生，以加快该领域发展的教材。该领域发展非常迅速，这本书提供了一条学习该领域基本思想和了解该领域现状的快捷之路。我认为该书内容丰富、刺激，相信读者也会有同样的感触。

Jim Gray
Microsoft Research

前　言

在过去的数十年中，我们产生和收集数据的能力已经迅速提高。起作用的因素包括条码在大部分商业产品中的广泛使用，许多商务、科学和行政事务的计算机化，以及由文本和图像扫描平台到卫星遥感系统的数据收集工具的进步。此外，作为全球信息系统的万维网的流行，已经将我们淹没在数据和信息的汪洋大海中。存储数据的爆炸性增长业已激起对新技术和自动工具的需求，以便帮助我们将海量数据转换成信息和知识。

本书考察数据挖掘的技术和概念。数据挖掘是数据库系统和新的数据库应用的一个有希望的、欣欣向荣的学科前沿。数据挖掘通常又称数据库中知识发现（KDD），是自动的或方便的模式提取，这些模式代表隐藏在大型数据库、数据仓库或其他大量信息存储中的知识。

数据挖掘是一个多学科领域，从多个学科汲取营养。这些学科包括数据库技术、人工智能、机器学习、神经网络、统计学、模式识别、知识库系统、知识获取、信息检索、高性能计算和数据可视化。我们从数据库角度提供本书中的材料。即是，我们集中讨论关于隐藏在大型数据库中的模式发现技术的可行性、有用性、有效性和可伸缩性问题。这样，本书不打算作为数据库系统、机器学习、统计学或其他某些领域的导论，尽管我们确实提供了这些领域必要的背景材料，以便读者理解它们各自在数据挖掘中的作用。本书是数据挖掘的全面介绍，与数据库主要问题一起讨论。对于计算科学的学生、应用开发者、商务人员以及涉及以上列举的学科的研究者，本书应当是有用的。

数据挖掘出现于20世纪80年代后期，90年代有了突飞猛进的发展，并可望在新千年继续繁荣。本书从数据库研究者的角度提供该领域的全面情况，介绍有趣的数据挖掘技术和系统，并讨论应用和研究方向。写作本书的重要动机是需要建立一个学习数据挖掘的有组织的框架——由于这个快速发展领域的多学科特点，这是一个具有挑战性的任务。我们希望本书有助于具有不同背景和经验的人交换关于数据挖掘的见解，为进一步促进这个令人激动的、不断发展的领域的成长做出贡献。

写给教师

本书旨在提供数据挖掘领域的一个广博的然而也是深入的概览。对于讲授高年级本科生或一年级研究生的数据挖掘课程，本书是有用的。此外，每章都包含了数据库或人工智能课程选题方面的材料。我们试图使得每章尽可能自包含，以便你不必顺序阅读每一章。对于本科生课程，可以使用第1～8章作为课程的核心材料。余下的课堂材料可以由第9、10章介绍的更高级的课题中选择。对于研究生课程，可以选择一学期讲完全书。

每章后面都有一些习题，适合作为课后作业。这些习题或者是短问题，用于测验对内容的掌握；或者是长问题，需要分析思考。

写给学生

我们希望这本教材将激发你对刚刚开始然而正在发展的数据挖掘领域的兴趣。我们试图以清晰的方式提供材料，仔细地解释所涵盖的课题。每一章的结尾有一个小结，介绍要点。全书包含了许多图和解释，以便使本书成为更令人愉快的和对“读者友好”。尽管本书作为教材编写，我们也试图组织它，使得它也是一本有用的参考书或手册，如果你今后决定在数据挖掘方面求职的话。

为阅读本书，你需要知道什么？

- 你应当具有一些关于数据库系统的概念和术语方面的知识。然而，我们确实试图提供数据库技术基础的足够背景，以便如果你的记忆有点生锈，你也能够理解本书的讨论。你应当具有一些数据库查询知识，尽管任何特定的查询语言知识不是必需的。
- 你应当具有一些程序设计经验。特别是，你应当能够阅读伪代码，以及理解简单的数据结构，如多维数组。
- 在统计学、机器学习或模式识别方面具有一些预备知识是有帮助的。然而，我们将使你从数据库角度熟悉这些领域与数据挖掘相关的基本概念。

写给专业人员

本书旨在涵盖数据挖掘领域的广泛课题。这样，本书是关于该主题的一本优秀手册。由于每一章的编写尽可能独立，你可以专注于你最感兴趣的课题。本书的大部分适合像你一样希望学习数据挖掘的关键思想的应用程序员和信息服务管理者。

所提供的技术和算法是实用的。本书介绍的算法适合于发现隐藏在大型的现实数据库中的模式，而不是挑选在小型“玩具”数据库上运行良好的算法。在第10章，我们简略地讨论了数据挖掘系统的商业应用，以及有希望的研究原型。本书提供的每个算法都用伪代码解释，但经过精心策划，使得不熟悉C或C++的程序员易于理解。如果你想实现算法，你会发现将我们的伪代码转换成选定的程序设计语言程序是一项直接了当的任务。

本书的组织

本书的组织如下：

第1章提供关于数据挖掘的多学科领域的导论。该章简略介绍数据库技术的发展，这些发展导致需要数据挖掘，以及数据挖掘潜在应用的重要性；描述数据挖掘系统的基本结构，简略介绍数据库系统和数据仓库系统的概念；根据挖掘的知识类型，介绍数据挖掘任务的详细分类；介绍数据挖掘系统的分类，并讨论该领域的主要挑战。

第2章是数据仓库和OLAP（联机分析处理）的引论。课题包括数据仓库和多维数据库，数据立方体结构，联机分析处理的实现，以及数据仓库和数据挖掘的关系。

第3章介绍挖掘之前的数据预处理技术。讨论数据清理、数据集成和转换、数据归约的方法，包括动态和静态离散化概念分层的使用。本章还介绍了概念分层的自动生成。

第4章介绍定义数据挖掘任务说明的数据挖掘原语。该章介绍数据挖掘查询语言（DMQL），

给出一些数据挖掘查询的例子。本章还讨论了其他语言，以及图形用户界面构造和数据挖掘系统结构。

第5章介绍概念描述技术，包括特征和区分。介绍面向属性的概化技术，以及它的不同实现，包括概化关系技术和多维数据方技术。解释多种形式的知识表示和可视化，讨论相关分析。给出多抽象层类比较方法和具有兴趣度度量的特征规则和区分规则的提取方法。此外，还讨论描述式挖掘的统计度量。

第6章介绍在事务数据库以及关系数据库和数据仓库中挖掘关联规则的方法。包括关联规则的分类，基本Apriori算法和它的变形，挖掘多层关联规则、多维关联规则、量化关联规则和相关规则的技术。介绍一种称作频繁模式增长的新技术，它挖掘频繁模式，而不产生候选项集。该章还讨论通过基于限制的挖掘和使用兴趣度度量对规则搜索聚焦，找出有趣规则的策略。

第7章介绍数据分类和预测方法，包括判定树归纳、贝叶斯分类、后向传播的神经网络技术、*k*-最邻近分类法、基于案例的推理、遗传算法、粗糙集理论和模糊集方法。介绍基于源自关联规则挖掘的概念分类，并介绍回归方法和讨论有关分类法精确度的问题。

第8章介绍聚类分析方法。首先介绍数据聚类概念，然后提供若干主要的数据聚类技术，包括基于划分的聚类、层次聚类和基于模型的聚类。介绍聚类连续数据、离散数据和多维数据立方体中的数据方法。详细讨论聚类算法的可伸缩性。

第9章讨论先进的数据库系统中的数据挖掘方法。包括在面向对象数据库、空间数据库、多媒体数据库、时间序列数据库、文本数据库和万维网中的数据挖掘。

最后，在第10章，我们总结本书提供的概念，并讨论数据挖掘的应用和一些具有挑战性的研究问题。

遍及全书，楷体字用于强调定义的术语，而黑体字用于突出主要思想。

错误

或许，本书可能包含打字错误、差错或遗漏。如果你发现错误，提出关于附加的习题的建议或者其他建设性批评，我们将乐于听到。我们欢迎并感谢你的建议。你可以将你的意见寄到

Data Mining:Concepts and Techniques
Intelligent Database Systems Research Laboratory
School of Computing Science
Simon Fraser University
Burnaby, British Columbia
Canada V5A 1S6
Fax: (604) 291-3045

此外，你也可以使用电子邮件提交查错报告，申请已知错误清单，或提出建设性意见。为收到指导，发一个e-mail到*dmbook@cs.sfu.ca*，消息头部写上“Subject: Help”。很抱歉，我们不能亲自回答所有的电子邮件消息。书中的错误和涉及本书的其他更新信息可以通过Web地址*www.cs.sfu.ca/~han/DM_Book*找到。

目　录

第1章 引 言

本书是一个导论，介绍什么是*数据挖掘*，什么是*数据库中知识发现*。书中的材料从数据库角度提供，特别强调发现隐藏在大型数据集中有趣数据模式的基本数据挖掘概念和技术。所讨论的实现方法主要是面向可伸缩的、有效的数据挖掘工具的开发。在本章中，我们将学习数据挖掘如何成为数据库技术自然演化的一部分，为什么数据挖掘是重要的，以及如何定义数据挖掘。我们将学习数据挖掘系统的一般结构，并考察挖掘的数据种类，可以发现的模式类型，以及什么样的模式提供有用的知识。除学习数据挖掘系统的分类之外，还将看到建立未来的数据挖掘工具所面临的挑战性问题。

1.1 什么激发了数据挖掘，为什么它是重要的

需要是发明之母。

近年来，数据挖掘引起了信息产业界的极大关注，其主要原因是存在大量数据，可以广泛使用，并且迫切需要将这些数据转换成有用的信息和知识。获取的信息和知识可以广泛用于各种应用，包括商务管理、生产控制、市场分析、工程设计和科学探索等。

数据挖掘是信息技术自然演化的结果。演化过程的见证是数据库业界开发以下功能（见图1-1）：*数据收集和数据库创建*，*数据管理*（包括数据存储和检索，数据库事务处理），以及*数据分析与理解*（涉及数据仓库和数据挖掘）。例如，数据收集和数据库创建机制的早期开发已成为稍后数据存储和检索、查询和事务处理有效机制开发的必备基础。随着提供查询和事务处理的大量数据库系统广泛付诸实践，数据分析和理解自然成为下一个目标。

自20世纪60年代以来，数据库和信息技术已经系统地从原始的文件处理演化到复杂的、功能强大的数据库系统。自70年代以来，数据库系统的研究和开发已经从层次和网状数据库系统发展到开发关系数据库系统（数据存放在关系表结构中；见1.3.1节）、数据建模工具、索引和数据组织技术。此外，用户通过查询语言、用户界面、优化的查询处理和事务管理，可以方便、灵活地访问数据。**联机事务处理**(OLTP)将查询看作只读事务，对于关系技术的发展和广泛地将关系技术作为大量数据的有效存储、检索和管理的主要工具作出了重要贡献。

自80年代中期以来，数据库技术的特点是广泛接受关系技术，研究和开发新的、功能强大的数据库系统。这些使用了先进的数据模型，如扩充关系模型、面向对象模型、对象-关系模型和演绎模型。包括空间的、时间的、多媒体的、主动的和科学的数据库、知识库、办公信息库在内的面向应用的数据库系统百花齐放。涉及分布性、多样性和数据共享问题被广泛研究。异种数据库和基于Internet的全球信息系统，如WWW也已出现，并成为信息产业的生力军。

在过去的30年中，计算机硬件稳定的、令人吃惊的进步导致了功能强大的计算机、数据收集设备和存储介质的大量供应。这些技术大大推动了数据库和信息产业的发展，使得大量数据库和信息存储用于事务管理、信息检索和数据分析。

现在，数据可以存放在不同类型的数据库中。最近出现的一种数据库结构是**数据仓库**（见1.3.2节）。这是一种多个异种数据源在单个站点以统一的模式组织的存储，以支持管理决策。数据仓库技术包括数据清理、数据集成和**联机分析处理（OLAP）**。OLAP是一种分析技术，具有汇总、合并和聚集功能，以及从不同的角度观察信息的能力。尽管OLAP工具支持多维分析和决策，对于深层次的分析，如数据分类、聚类和数据随时间变化的特征，仍然需要其他分析工具。

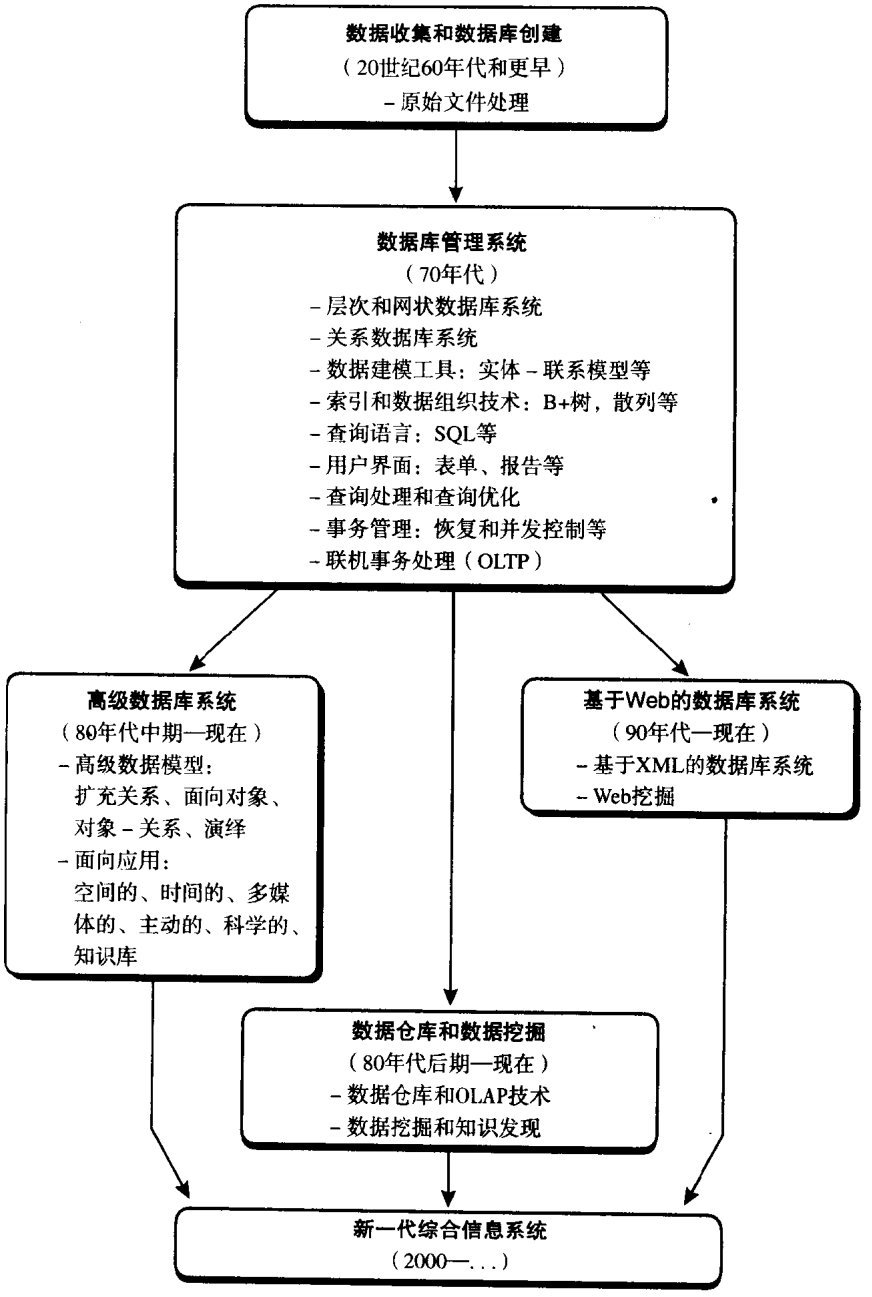

图1-1　数据库技术的演化